

SEGMENTASI PELANGGAN MAJALAH PADA SITUS WEB E-COMMERCE DENGAN K-MEANS++ DAN METODE RFM

Andrew Lomaksan Manuel Tampubolon^{*1}, Thio Marta Elisa Yuridis Butar Butar², Siti Rochimah³

^{1,2,3}Institut Teknologi Sepuluh Nopember, Surabaya

Email: ¹6025231025@student.its.ac.id, ²6025231009@student.its.ac.id, ³siti@if.its.ac.id

^{*}Penulis Korespondensi

(Naskah masuk: dd mmm yyyy, diterima untuk diterbitkan: dd mmm yyyy)

Abstrak

Segmentasi pelanggan merupakan salah satu metode yang dapat diterapkan untuk memaksimalkan peluang bisnis agar tetap kompetitif dalam persaingan pasar. *Artificial Intelligence* (AI) dapat meningkatkan pemahaman tentang cara menyegmentasikan pelanggan berdasarkan riwayat transaksi. Penelitian ini menerapkan metode *Recency*, *Frequency*, and *Monetary* (RFM) yang dipadukan dengan algoritma K-Means++ untuk melakukan segmentasi pelanggan dengan memperhatikan nilai k yang paling optimal untuk menentukan jumlah *cluster*. Kerangka kerja CRISP-DM yang digunakan dalam makalah ini membantu mempertahankan proses yang konsisten. Menggunakan pendekatan statistik sederhana untuk mengklasifikasikan setiap fitur dalam RFM dengan label rendah, sedang, dan tinggi, model ini dapat menangkap pola untuk menyegmentasikan pelanggan. Model ini dibangun dengan nilai $k = 3$; hasilnya, nilai WSS sebesar 843,214747, dan *silhouette score* sebesar 0,638181. *Cluster 0* memiliki nilai RFM rata-rata sebesar 1,14 (rendah), 1,20 (rendah), dan 301.640 (rendah). *Cluster 1* memiliki nilai RFM rata-rata sebesar 249,61 (tinggi), 2,62 (sedang), dan 799,934 (sedang). *Cluster 2* memiliki nilai RFM rata-rata sebesar 233,01 (sedang), 6,41 (tinggi), dan 2018,088 (tinggi).

Kata kunci: Segmentasi Pelanggan, E-Commerce, K-Means++, RFM, CRISP-DM.

SEGMENTATION OF MAGAZINE SUBSCRIBERS ON E-COMMERCE WEBSITE USING K-MEANS++ AND RFM METHOD

Abstract

Customer segmentation is one method that can be applied to maximize business opportunities to remain competitive in market competition. *Artificial Intelligence* (AI) can increase understanding of how to segment customers based on their history of transactions. This paper applies the *Recency*, *Frequency*, and *Monetary* (RFM) method combined with the K-Means++ algorithm to segment customers by paying attention to the most optimal k value to determine the number of clusters. The CRISP-DM framework used in this paper helps maintain a consistent process. Using a simple statistical approach to classify each feature in an RFM with low, medium, and high labels, the model can capture patterns to segment customers. The model is built with a value of $k = 3$; as a result, the WSS value is 843.214747, and the *silhouette score* is 0.638181. *Cluster 0* has an average RFM value of 1.14 (low), 1.20 (low), and 301,640 (low), respectively. *Cluster 1* had average RFM values of 249.61 (high), 2.62 (medium), and 799.934 (medium), respectively. *Cluster 2* had average RFM values of 233.01 (medium), 6.41 (high), and 2018.088 (high), respectively.

Keywords: Customer Segmentation, E-Commerce, K-Means++, RFM, CRISP-DM.

1. PENDAHULUAN

E-commerce adalah platform digital yang memfasilitasi transaksi jual beli online antara penjual dan konsumen. Tujuan utama *E-commerce* adalah menjual produk dalam bentuk barang fisik dan layanan digital secara online (Ros'ario & Raimundo, 2021). Tujuan utama *E-commerce* adalah untuk menjual produk dalam bentuk barang fisik dan layanan digital secara online (Jain, Malviya & Arya, 2021). Platform *E-commerce* menawarkan berbagai

fitur dan kemudahan, seperti katalog produk dan keranjang belanja. Platform *E-commerce* juga menyediakan berbagai opsi pembayaran bagi pelanggan. Fitur tambahan memungkinkan pelanggan untuk menemukan, memilih, dan membeli produk atau layanan. Mengenai struktur bisnis, *E-commerce* mencakup beragam model pemasaran. Ini termasuk pengecer online, layanan berlangganan, pasar digital, dan entitas B2B yang mengkhususkan

diri dalam menjual produk atau layanan ke bisnis lain (Chen et al., 2022).

Strategi bisnis yang efektif diperlukan untuk memaksimalkan peluang yang ada. Mengembangkan strategi komprehensif yang mencakup segmentasi pasar dan alokasi sumber daya praktis menjadi sangat penting demi meningkatkan angka penjualan produk. Selain itu, bisnis perlu untuk memperoleh keunggulan kompetitif di pasar (Dwivedi et al., 2021). Sebagai salah satu solusi, kecerdasan buatan dapat memberi bisnis pemahaman tentang karakteristik dan perilaku pelanggan. Ini juga membantu bisnis mendapatkan pemahaman yang lebih dalam tentang perilaku konsumen (Fauzan & Davin, 2023; Verma et al., 2021). Dengan menganalisis data, kecerdasan buatan dapat membantu bisnis mengidentifikasi tren dalam aktivitas pengguna. Hal ini juga dapat membantu mendapatkan wawasan tentang preferensi dan karakteristik pelanggan. Kecerdasan buatan cukup untuk memberikan rekomendasi yang dipersonalisasi untuk meningkatkan pengalaman pelanggan dalam berbelanja (Verma et al., 2021; Haleem et al., 2022).

Analisis dengan kecerdasan buatan sangat penting dengan memungkinkan organisasi berbasis data. Analisis data dapat membantu pengambilan keputusan yang lebih cepat dan mendapatkan wawasan tentang karakteristik pelanggan (Verma et al., 2021; Haleem et al., 2022). Kecerdasan buatan dapat menambah wawasan konsumen yang lebih dalam tentang manfaat produk yang dipasarkan. Selain itu pelanggan dapat lebih memahami cara mengkategorikan dan mendorong pelanggan ke langkah berikutnya dalam proses belanja mereka (Fauzan & Davin, 2023; Verma et al., 2021). Kecerdasan buatan menawarkan wawasan kepada konsumen tentang perilaku konsumen yang penting untuk daya tarik pelanggan (Verma et al., 2021; Jain, Wadhvani, & Eastman, 2023). Menerapkan kecerdasan buatan untuk penjualan yang efektif membutuhkan strategi yang komprehensif. Proses ini mencakup strategi data, rekayasa data, tata kelola, manajemen perubahan, dan budaya (Haleem et al., 2022). Menurut laporan AI: Built to Scale dari Accenture, 84 persen eksekutif bisnis percaya bahwa kecerdasan buatan dapat membantu mencapai tujuan pertumbuhan mereka. Namun, 76 persen mengakui membutuhkan bantuan untuk meningkatkan kecerdasan buatan di seluruh bisnis mereka (Fauzan & Davin, 2023). Oleh karena itu, bisnis harus mengambil langkah-langkah yang diperlukan untuk mengembangkan strategi yang komprehensif. Kecerdasan buatan perlu untuk dimanfaatkan dengan baik agar tetap terdepan dan kompetitif di pasar.

Kecerdasan buatan diperlukan dalam mempersonalisasi dan tetap kompetitif di pasar. Algoritma kecerdasan buatan dapat digunakan untuk menyegmentasikan pelanggan. Algoritma *unsupervised machine learning* adalah algoritma pembelajaran mesin tanpa pengawasan untuk

menemukan kelompok titik data yang terkait erat (Verma et al., 2021; Sarker, 2021). Algoritma pengelompokan K-Means adalah salah satu contoh algoritma pembelajaran mesin standar untuk *clustering* data. Algoritma tersebut baik untuk kasus segmentasi pelanggan (Siebert et al., 2018). Segmentasi pelanggan adalah proses mengklasifikasikan pelanggan dengan karakteristik serupa ke dalam segmen serupa (Sarker, 2022). Algoritma pengelompokan membantu untuk lebih memahami karakteristik pelanggan, baik dalam hal demografi dan perilaku dinamis pelanggan (Verma et al., 2021). Metode ini dapat dikombinasikan dengan metode lain seperti Recency, Frequency, dan Monetary (RFM). Metode RFM membantu mengevaluasi pelanggan berdasarkan perilaku belanja mereka (Christy et al., 2021). Metode ini juga dapat digunakan untuk meningkatkan analisis pelanggan dan prediksi pengalaman berbelanja (Khajvand et al., 2011). Informasi mengenai bagaimana pelanggan berbelanja dapat dikumpulkan dengan menganalisis riwayat transaksi pelanggan dengan memanfaatkan data transaksi yang tersimpan dalam *database*.

Christy et al. mengelompokkan pelanggan menjadi beberapa kelompok berdasarkan nilai Recurrence, Frekuensi, dan Moneter (RFM) seperti yang ditekankan dalam penelitian ini. Algoritma pengelompokan seperti K-means dan fuzzy C-means digunakan untuk menganalisis perilaku konsumen menggunakan data transaksi jual beli. Pemilihan *centroid* awal mampu mengurangi jumlah iterasi dan waktu yang dibutuhkan untuk segmentasi pelanggan. Hasilnya akan dibandingkan dengan metode pengelompokan konvensional dengan mempertimbangkan kekompakan kluster, waktu eksekusi, dan jumlah iterasi. Hasil studi menunjukkan bahwa *clustering* membantu perusahaan memberikan rekomendasi produk, mengidentifikasi tren, dan menyesuaikan program pemasaran (Christy et al., 2021).

Jinfeng Zhou dkk. Mendemonstrasikan model RFMT (Recency, Frequency, Monetary, and Interpurchase Time) untuk menyegmentasikan pelanggan di industri ritel. Model ini dapat digunakan untuk mengidentifikasi berbagai kelompok pelanggan berdasarkan perilaku pembelian mereka. Hasil segmentasi dapat memberikan rekomendasi untuk strategi bisnis, seperti mengalokasikan sumber daya pemasaran dan rekomendasi produk yang disesuaikan. Namun, perlu ada lebih banyak informasi tentang metode yang digunakan untuk mengumpulkan data pelanggan. Selain itu, makalah ini juga perlu memberikan informasi tentang ukuran sampel yang digunakan dalam penelitian. Hasil penelitian dapat menunjukkan informasi yang dapat membantu mengevaluasi reliabilitas dan generalisasi hasil segmentasi (Zhou, Wei & Xu, 2021).

Anitha et al. menerapkan konsep *business intelligence* dalam mengidentifikasi pelanggan

potensial dengan menyediakan data yang relevan dan tepat waktu untuk industri ritel. Studi ini menganalisis riwayat transaksi untuk mengidentifikasi perilaku belanja konsumen dan keuntungan bisnis. Model RFM (Recency, Frequency, dan Monetary) dikombinasikan dengan algoritma K-Means untuk menyegmentasikan pelanggan. Analisis RFM menghitung kebaruan, frekuensi, dan nilai moneter setiap pelanggan untuk proses segmentasi. Algoritma K-Means memilih jumlah cluster berdasarkan *silhouette score* berdasarkan nilai RFM (Anitha & Patil, 2022).

Rahim et al. menggunakan model Recency, Frequency, and Monetary (RFM) yang dikombinasikan dengan model pembelajaran mesin seperti Multi-Layer Perceptron (MLP) dan Support Vector Machine (SVM) untuk mengklasifikasikan pelanggan. Penelitian ini bertujuan untuk mengklasifikasikan pelanggan berdasarkan perilaku belanja di industri ritel (Rahim et al., 2021).

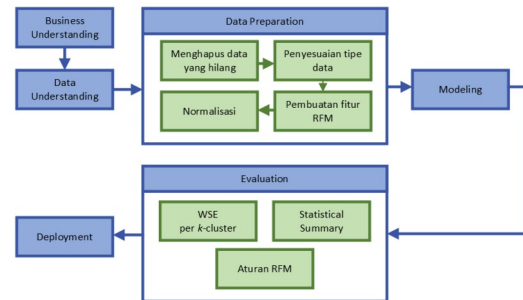
Dedi et al. mempelajari segmentasi pelanggan berdasarkan nilai Recency, Frequency, and Monetary (RFM) dengan K-means tradisional dan Fuzzy C-means. Data yang digunakan dalam penelitian adalah data transaksional untuk menganalisis perilaku pelanggan. Makalah ini mengusulkan metode baru dengan memilih *centroid* awal di K-Means dan bertujuan untuk menyegmentasikan pelanggan dengan iterasi dan waktu yang berkurang. Algoritma RM K-Means yang diusulkan menghabiskan lebih sedikit waktu dan mengurangi iterasi, membuatnya lebih efektif. Segmentasi berfokus pada kebaruan, frekuensi, dan nilai moneter, memungkinkan perusahaan untuk menyesuaikan strategi pemasaran berdasarkan perilaku pembelian (Dedi et al., 2019).

Berdasarkan penelitian sebelumnya tentang pengelompokan pelanggan pada *platform E-commerce*, metode klasifikasi dengan algoritma K-Means *clustering* yang dikombinasikan dengan metode RFM diusulkan untuk menemukan pola perilaku pelanggan dalam berbelanja. Metode ini dikombinasikan dengan metode CRISP-DM untuk membakukan proses pemrosesan data ke hasil yang lebih representatif. Kami menggunakan *silhouette score* untuk memutuskan k apa parameter yang paling optimal.

2. METODOLOGI PENELITIAN

Penelitian menggunakan metode CRISP-DM yang terdiri dari 5 langkah mulai dari *business understanding* hingga *evaluation*. Prosesnya meliputi *business understanding*, *data understanding*, *modeling*, *evaluation*, and *deployment* (Wirth & Hipp, 2000). Namun proses *deployment* tidak dilakukan pada penelitian ini. Proses yang diterapkan pada penelitian terbatas pada *evaluasi*, yaitu dihasilkan cluster yang representatif untuk mensegmentai pelanggan. Selain itu pada beberapa proses CRISP-DM berisi sub-proses lain seperti *preparation* and *evaluation* data, seperti yang

ditunjukkan pada gambar 1. Berikut adalah proses yang diterapkan dalam penelitian:



Gambar 1. Alur penelitian berdasarkan metode CRISP-DM

2.1. Business Understanding

Situs E-commerce menyediakan berbagai produk majalah cetak dan majalah elektronik untuk segmen usia yang berbeda dengan segment kelompok anak-anak, pria, dan wanita. Proses yang terjadi dalam E-commerce adalah serangkaian proses dari pelanggan mendaftar, memasukkan barang ke dalam keranjang, *checkout*, dan pembayaran hingga transaksi selesai, baik itu transaksi berhasil, kedaluarsa, atau dibatalkan. Sebagai aspek yang dapat menarik pelanggan, E-commerce melakukan promosi seperti diskon produk, bundling produk, bonus hadiah, dan voucher diskon. Namun, saat ini, promosi belum dilakukan sesuai target pelanggan. Hal ini diperlukan untuk merencanakan proses promosi kepada pelanggan potensial dengan tujuan tidak hanya meningkatkan pendapatan, Hal ini juga diharapkan untuk menawarkan berbagai promosi yang menarik pelanggan.

2.2. Data Understanding

Data dalam E-commerce terdiri dari transaksi pelanggan terhadap *e-magazine* dan produk cetak. Transaksi dimulai dari 1 Maret 2021 hingga 1 Maret 2023. Kriteria data riwayat transaksi hanya transaksi berhasil yang terdiri dari data 5847 baris dan 11 atribut seperti tanggal transaksi, tanggal pembuatan akun, kode transaksi, email, id pengguna, merek, harga berlangganan, jumlah item, harga satuan, jenis produk, dan jenis langganan. Data diperoleh dari *database* dengan SQL (Structured Query Language). Secara umum, data yang tersedia mewakili komponen yang diperlukan dalam metode RFM. Semua atribut ini nantinya akan digunakan dalam proses pra pemrosesan data.

2.3. Data Preparation

Data yang disimpan dalam database E-commerce belum mengalami pemrosesan seperti pada data yang disimpan di *data warehouse*. Bentuk data seperti tipe data, format, dan ukuran, dan menghapus nilai *null* menjadi perhatian dalam proses

data mining dalam data transaksi E-commerce. Dalam proses ini akan disesuaikan data dengan tipe data yang sesuai untuk diproses dengan benar. Kemudian, data rahasia akan disamarkan terlebih dahulu, seperti *trans_id* dan *user_id*. Atribut baru dibentuk dengan nama produk, yang merupakan kombinasi dari merek, durasi berlangganan, dan jenis produk (E-magazine atau cetak), seperti pada deskripsi pada tabel 1.

Tabel 1. Atribut dataset

Nama Atribut	Deskripsi
<i>trans_date</i>	Tanggal transaksi
<i>acc_created_date</i>	Tanggal pembuatan akun
<i>trans_id</i>	Nomor transaksi
<i>user_id</i>	Id unik transaksi
<i>brand</i>	Merk produk
<i>product_name</i>	Nama produk dan lama langganan
<i>quantity</i>	Jumlah item transaksi
<i>sub_total</i>	Total pembayaran per produk

Penelitian ini menggunakan data dan data numerik, sedangkan data kategoris seperti *user_id* digunakan sebagai *unique identifier*. Fitur lain dari himpunan data yang tidak digunakan akan dihilangkan. Untuk menghasilkan fitur baru yang dapat diproses dengan metode RFM, metode analisis berdasarkan data riwayat transaksional untuk pemasaran (Hughes, 1994). *Recency* adalah jarak antara pembelian awal dan pembelian berikutnya. *Frequency* adalah perhitungan berapa kali pelanggan berbelanja dalam periode yang telah ditentukan. *Monetary* adalah hasil perhitungan total transaksi nasabah dalam periode yang telah ditentukan.

Jika pelanggan memiliki beberapa transaksi, nilai *recency* didasarkan pada tanggal dan frekuensi transaksi terakhir. Sedangkan *monetary* adalah nilai total dari keseluruhan pengeluaran. Semua fitur pertama-tama akan dinormalisasi dengan *StandardScaler* untuk menormalkan data sehingga meminimalkan nilai kesalahan (Thara et al., 2019). Kemudian, data tersebut akan diolah dengan reduksi dimensi dengan *Principal Component Analysis* (PCA) menjadi data dua dimensi.

2.4. Modeling

Model *machine learning* yang digunakan adalah proses *clustering* dengan menggunakan K-Means++ untuk mengatasi permasalahan algoritma K-Means klasik. Bentuk K-Means klasik secara acak menginisialisasi *centroid* awal, menghasilkan hasil pengelompokan yang berbeda yang dapat menyebabkan kesalahan atau memperlambat algoritma untuk mencapai konvergensi (Gao et al., 2021). Iterasi yang ditetapkan akan membatasi K-Means ++ hingga 300 kali hingga model mencapai konvergensinya. Model optimal dihasilkan dari pemilihan parameter optimal nilai *k*. Untuk proses evaluasi, parameter *n_init* digunakan sebagai parameter optimal ditentukan menggunakan

visualisasi nilai WSS sehingga dihasilkan grafik untuk evaluasi, kemudian dengan *elbow method* dan perhitungan berdasarkan *silhouette score* terbaik (Rousseeuw, 1987). Algoritma K-Means dengan parameter terbaik nantinya akan digunakan untuk mengelompokkan data pelanggan.

2.5. Evaluation

Dengan menetapkan nilai *k*, cluster akan dibentuk menggunakan data yang telah melalui pra-pemrosesan dan *dimension reduction* PCA. Data dikelompokkan menurut *cluster* masing-masing, di mana setiap pengguna akan diberi label kelas untuk divisualisasi. Data dikelompokkan berdasarkan kategori. Setiap *cluster* dapat dikelompokkan menjadi rendah, sedang, dan tinggi berdasarkan nilai rata-rata setiap *cluster*. Pertama, hitung nilai kuartil 1 dari setiap fitur himpunan data. Kedua, hitung nilai kuartil 3 dari setiap fitur himpunan data. Ketiga, data yang nilai rata-ratanya di bawah kuartil satu dikategorikan sebagai 'rendah'. Keempat, data yang nilai rata-ratanya berada di kisaran kuartil 1 hingga kuartil tiga dikategorikan sebagai 'sedang'. Kelima, data yang nilai rata-ratanya di atas kuartil tiga dikategorikan 'tinggi'.

3. HASIL

3.1. Bentuk Data

Terdapat 5.847 baris dan 9 kolom (fitur) data riwayat transaksi sebelum proses pra-pemrosesan. Data tidak memiliki nilai null karena data transaksi yang diambil dengan SQL relatif bersih. Namun data masih perlu dilakukan pra-pemrosesan untuk beberapa fitur yang berbentuk kategorik dan numerik. Fitur *user_id* tetap dipertahankan sebagai *unique identifier* untuk setiap transaksi. Setelah proses pra-pemrosesan, 3900 baris data dihasilkan fitur baru dalam bentuk fitur *monetary*, *frequency*, dan *recency* dalam bentuk numerik. Metode normalisasi data numerik menghasilkan distribusi normal dengan metode *standard scaler*. Rumus *standard scaler* dapat dijelaskan seperti rumus berikut (Thara et al., 2019):

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

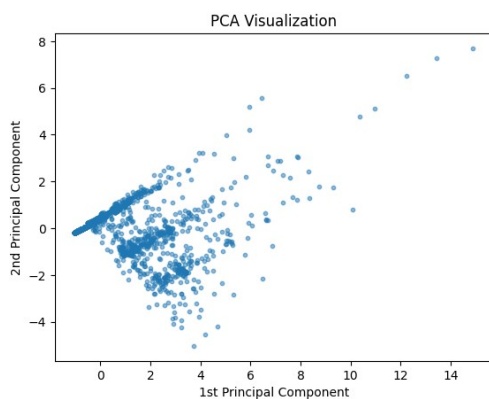
di mana *x* adalah nilai awal, μ adalah mean, dan σ adalah standar deviasi. Normalisasi ini bertujuan untuk mengubah nilai rata-rata distribusi menjadi nol dan standar deviasinya menjadi 1. Dengan demikian, data yang diperoleh memiliki distribusi yang seragam sehingga data dapat digunakan untuk proses clustering. Beberapa sampel data dapat dilihat pada tabel 2.

Tabel 2. Fitur setelah proses normalisasi

user_id	recency	frequency	monetary
001b5eddf7	0.4387	0.4416	-0.2096
0023cedbf9	-0.9347	-0.4402	-0.2096
0024ae4898	0.0752	0.4416	-0.2096
002d8b5849	-0.4700	-0.4402	-0.2096

3.2. Reduksi Dimensi

PCA adalah metode untuk mengekstraksi pola dalam fitur dengan mengurangi dimensi. PCA menggabungkan beberapa variabel menjadi dua atau tiga komponen. PCA bekerja dengan mengurangi variasi data dan menemukan pola yang kuat dari kumpulan data (Wold, Esbensen & Geladi, 1987). Fitur *monetary*, *recency*, dan *frequency* yang sebelumnya dinormalisasi akan dikurangi dari 3 (*monetary*, *recency*, dan *frequency*. *user_id* tidak termasuk) dimensi menjadi dua ukuran. Pemilihan dua dimensi berikut bertujuan untuk meminimalisir dimensi komponen dan proses visualisasi data dengan *scatter plot*. Pada gambar 2, dapat dilihat bahwa data semakin berhimpit tetapi belum ada pengelompokan yang jelas.



Gambar 2. PCA 2 dimensi sebelum proses *clustering*

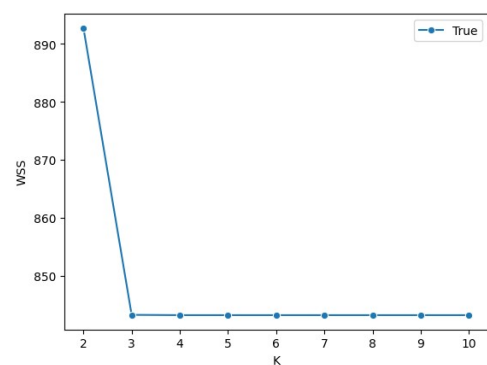
3.3. Analisis Jumlah Cluster

Sebelum proses *clustering*, pertama dilakukan proses analisis cluster. Proses analisis *cluster* bertujuan untuk menemukan jumlah *cluster* atau nilai *k* yang optimal. Parameter *k* adalah input yang digunakan sebagai nilai parameter fungsi K-Means pada *library* sklearn. Analisis cluster menggunakan *elbow method* dan *silhouette score* yang telah dilakukan pada tahap sebelumnya. Nilai (Dalam Jumlah Kuadrat) WSS dihasilkan dengan melakukan iterasi sebanyak sepuluh kali untuk setiap *cluster*. Pada gambar 3, dapat dilihat bahwa memplot nilai WSS untuk setiap kelompok menghasilkan *elbow method*. Dengan begitu, kita dapat menentukan bahwa nilai *k* terbaik berdasarkan *elbow method* adalah *k* dengan nilai 3. Mendapatkan *silhouette score* terhadap semua transaksi *k* dari satu hingga

sepuluh dapat dilakukan dengan menggunakan rumus 2 berikut (Rousseeuw, 1987):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2)$$

di mana $a(i)$ adalah rata-rata perbedaan/jarak ke semua titik objek dan $b(i)$ adalah nilai minimum antara jarak rata-rata sampel dari sampel ke *cluster* lain. Dengan menghitung *silhouette score* (misalnya, *k* dalam K-Means), *silhouette score* tertinggi menunjukkan jumlah *cluster* terbaik.



Gambar 3. *Elbow method*

Kemudian dilakukan proses analisis *cluster* dengan menggunakan metode *silhouette score*. Iterasi sepuluh kali dilakukan untuk mendapatkan nilai *k* yang sesuai. Mengacu pada tabel 3, dapat disimpulkan bahwa nilai *k* yang sesuai berdasarkan *silhouette score* adalah *k* = 3, dengan nilai 0,63818. *Silhouette score* terbaik adalah yang paling signifikan dibandingkan dengan jumlah *cluster* lainnya.

Tabel 3. *Silhouette score*

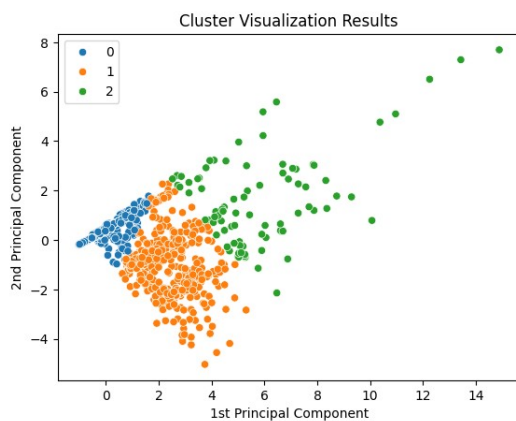
k	WSS	Silhouette Score
3	843.214747	0.638181
4	843.165970	0.638099
5	843.165970	0.638099
6	843.165970	0.638099
7	843.165970	0.638099
8	843.165970	0.638099
9	843.165970	0.638099
10	843.165970	0.638099
2	892.717979	0.637027

3.4. Clustering

Clustering dilakukan dengan menggunakan metode K-Means++ untuk mendapatkan pola segmentasi pelanggan. Parameter yang diterapkan pada fungsi K-Means ++ adalah *n_init* = 3 nilai *k*). Selain itu, iterasi maksimum ditentukan dengan menentukan *max_iter* = 300. Ini bertujuan untuk menentukan batas iterasi hingga mencapai

konvergensi. Proses pembentukan model akan membagi data menjadi tiga kelompok yang berbeda. Model K-Means akan menghasilkan label untuk setiap *record data*.

Untuk itu, fitur label akan dibentuk untuk menginisiasi data yang disertakan dalam nomor *cluster*. gambar 4 menunjukkan bagaimana *cluster* didistribusikan berdasarkan pola data. Dengan membagi *cluster* menjadi tiga, maka akan terbentuk 3 *centroid*. Titik data akan dikelompokkan berdasarkan jarak dengan *centroid* terdekat. Hanya cluster yang terbentuk untuk saat ini; Kemudian, akan dilakukan proses untuk menemukan pola dari masing-masing *cluster*.



Gambar 4. Plotting data untuk tiap cluster

3.5. Pola untuk tiap cluster

Menentukan pola masing-masing cluster dilakukan dengan mencari nilai mean dari masing-masing *cluster* berdasarkan fitur. Kemudian nilai tersebut akan menjadi patokan untuk menentukan apakah *cluster* dengan fitur RFM-nya dapat dikategorikan sebagai 'low', 'medium', atau 'high'. Setiap fitur dihitung terlebih dahulu. Dikategorikan sebagai 'low' ketika di bawah Q1, 'medium' antara Q1 dan Q3, dan 'high' ketika lebih dari Q3.

Tabel 4 menunjukkan bagaimana pola data didasarkan pada nilai rata-rata untuk setiap *cluster* terhadap fitur-fiturnya. Nilai *recency* dapat dikategorikan sebagai 'rendah' berarti jarak antara satu pembelian dan pembelian berikutnya cenderung cepat. Nilai *recency* 'tinggi' berarti bahwa waktu antara satu pembelian dan pembelian berikutnya cenderung lambat. Nilai *frequency* dikategorikan sebagai 'rendah' jika Anda jarang melakukan pembelian, *frequency* 'tinggi' berarti pembelian yang sering. *Monetary* dikategorikan 'rendah' jika nilai uang yang dikeluarkan untuk melakukan pembelian cenderung rendah, sedangkan 'tinggi' berarti nilai pembelian tinggi. Nasabah yang diinginkan adalah nasabah yang memiliki nilai *recency* rendah, *frequency* tinggi, dan *monetary* tinggi. Berdasarkan

proses kategorisasi yang telah dijelaskan, karakteristik masing-masing cluster dapat digambarkan sebagai berikut:

Tabel 4. Karakteristik RFM untuk tiap cluster

Cluster	recency	frequency	monetary
0	low (1.14)	low (1.20)	low (301640)
1	high (249.61)	medium (2.62)	medium (799934)
2	medium (233.01)	high (6.41)	high (2018088)

pada data yang belum dinormalisasi, pola-pola tersebut dengan nilai mean masing-masing *cluster* untuk setiap fitur RFM dapat dijelaskan pada Tabel 5 sebagai berikut:

Tabel 5. Segmentasi RFM pada data riil

Segment	recency	frequency	monetary
Low	< 117.07	< 1.91	< 550787
	≥ 117.07	≥ 1.91	≥ 550787
Medium	dan	dan	dan
	≤ 249.61	≤ 6.41	≤ 2018088
High	≥ 249.61	≥ 6.41	≥ 2018088

Pelanggan pada *cluster* ini berjumlah 3303 pelanggan dari total 3900 pelanggan yang dianalisis. Jumlah ini sekitar 84,69% dari total pelanggan yang dianalisis atau mayoritas pembeli di E-commerce. Ini menunjukkan karakteristik pelanggan pada *cluster* 0. Nilai pembelian rata-rata pada *cluster* ini adalah Rp. 301.640, Hal ini dapat diklasifikasikan sebagai nilai pembelian yang rendah. Frekuensi rata-rata pembelian adalah 1,20 kali. Ini diklasifikasikan sebagai frekuensi pembelian yang rendah. Nilai rata-rata waktu yang cukup cepat hingga pembelian berikutnya adalah 1,14 hari. Berdasarkan nilai-nilai tersebut, dapat dikatakan bahwa pelanggan di *cluster* ini cenderung melakukan pembelian dalam waktu singkat, bahkan dengan nilai pembelian yang rendah.

Pelanggan pada *cluster* ini berjumlah 511 pelanggan dari total 3900 pelanggan yang dianalisis. Jumlah ini sekitar 13,10% dari total pelanggan yang dianalisis. Rata-rata nilai pembelian di *cluster* ini adalah Rp. 799.934 atau tergolong nilai pembelian menengah. Rata-rata frekuensi pembelian adalah 2,62 kali dan rata-rata nilai waktu hingga pembelian berikutnya cukup lama, yaitu 249,61 hari. Berbeda dengan pembeli pada *cluster* sebelumnya, karakteristik pembeli pada *cluster* ini membeli produk dengan nilai sedang namun akan kembali membeli produk dalam jangka waktu yang lama.

Pelanggan pada *cluster* ini berjumlah 86 pelanggan dari total 3900 pelanggan yang dianalisis. Jumlah ini sekitar 2,20% dari total pelanggan yang dianalisis. Rata-rata nilai pembelian di *cluster* ini yaitu Rp. 2.018.088 atau dengan nilai yang besar dengan frekuensi pembelian rata-rata 6,41 kali, dan

nilai rata-rata waktu hingga pembelian berikutnya cukup lama, yaitu 233,01 hari. Karakter pelanggan berikut ini unik karena frekuensinya hingga 3 kali lipat dari cluster lainnya, bahkan dengan nilai beli yang besar. Hanya saja jangka waktu pembeliannya cukup lama untuk pembelian berikutnya.

4. KESIMPULAN DAN SARAN

Proses segmentasi pelanggan E-commerce dapat dilakukan dengan menggunakan metode RFM. Dengan menerapkan algoritma K-Means ++, terbentuk tiga *cluster* berbeda dengan karakteristiknya masing-masing dihasilkan. *Clustering* dengan nilai $k = 3$ menghasilkan nilai WSS 843,214747 dan *silhouette scor* 0,638181 sebagai k terbaik. Dengan membagi *cluster* menjadi tiga, didapatkan pola-pola pembelian dengan mensegmentasi pelanggan. *Cluster* 0 memiliki nilai RFM rata-rata masing-masing 1,14 (rendah), 1,20 (rendah), dan 301.640 (rendah). *Cluster* 1 memiliki nilai RFM rata-rata masing-masing 249,61 (tinggi), 2,62 (sedang), dan 799,934 (sedang). *Cluster* 2 memiliki nilai RFM rata-rata masing-masing 233,01 (sedang), 6,41 (tinggi), dan 2018,088 (tinggi). Segmentasi dan karakteristik segmen itu sendiri cukup representatif untuk menjelaskan bagaimana setiap *cluster* dipolakan. Dengan begitu, bisnis seperti penjualan majalah E-commerce mendapatkan wawasan tentang perilaku belanja pelanggan. Hasilnya dapat digunakan untuk keputusan dalam berbagai proses, termasuk promosi.

Mengingat penelitian ini menggunakan perhitungan *range* untuk menentukan kategori RFM, kedepannya dapat ditentukan dengan menggabungkan metode yang telah diterapkan dengan cara lain, seperti *association rule mining* (Chen & Gunawan, 2023). Akan menarik jika proses segmentasi pelanggan menggunakan algoritma yang berbeda, seperti K-Medoids dan DBSCAN, untuk mengetahui seberapa baik masing-masing algoritma melakukannya dengan menghitung nilai DBI atau *silhouette score* (Aryuni, Madyatmadja & Miranda, 2018).

DAFTAR PUSTAKA

- ANITHA, P., PATIL, M. M. 2022. RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, 34(5), 1785–1792.
- ARYUNI, M., MADYATMADJA, E. D., MIRANDA, E. 2018, September. Customer segmentation in XYZ bank using K-means and K-medoids clustering. In 2018 International Conference on Information Management and Technology (ICIMTech) (pp. 412-416). IEEE.
- CHEN, A. H.-L., GUNAWAN, S. 2023. Enhancing Retail Transactions: A Data-Driven Recommendation Using Modified RFM Analysis and Association Rules Mining. *Applied Sciences*, 13(18), 10057
- CHEN, Y., LI, M., SONG, J., MA, X., JIANG, Y., WU, S., CHEN, G. L. 2022. A study of cross-border E-commerce research trends: Based on knowledge mapping and literature analysis. *Frontiers in Psychology*, 13.
- CHRISTY, A. J., UMAMAKESWARI, A., PRIYATHARSINI, L., NEYAA, A. 2021. RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences*, 33(10), 1251–1257.
- DEDI, M. I. DZULHAQ, K. W. SARI, S. RAMDHAN, R. TULLAH, SUTARMAN. Customer Segmentation Based on RFM Value Using K-Means Algorithm. 2019. Fourth International Conference on Informatics and Computing (ICIC), Semarang, Indonesia, 2019, pp. 1-7.
- THARA D.K., B.G, PREMASUDHA., XIONG, F. 2019. Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. *Pattern Recognition Letters*, 128, 544–550.
- DWIVEDI, Y. K., ISMAGILOVA, E., HUGHES, D. L., CARLSON, J., FILIERI, R., JACOBSON, J., JAIN, V., KARJALUOTO, H., KEFI, H., KRISHEN, A. S., KUMAR, V., RAHMAN, M. M., RAMAN, R., RAUSCHNABEL, P. A., ROWLEY, J., SALO, J., TRAN, G. A., WANG, Y. 2021. Setting the future of digital and social media marketing research: Perspectives and research propositions. *International Journal of Information Management*, 59, 102168.
- FAUZAN M., DAVIN. 2023. Unlocking Digital Business Success: Leveraging Artificial Intelligence in Social Media Analytics for Enhanced Customer Insights and Engagement. 10.13140/RG.2.2.33588.63361.
- HALEEM, A., JAVAID, M., ASIM QADRI, M., PRATAP SINGH, R., SUMAN, R. 2022. Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*, 3, 119–132.

HUGHES. 1994 op. cit.

GAO, M., PAN, S., CHEN, S., LI, Y., PAN, N., PAN, D., SHEN, X. 2021. Identification Method of Electrical Load for Electrical Appliances Based on K-Means ++ and GCN. *IEEE Access*, 9, 27026–27037.

JAIN, V., MALVIYA, B., ARYA, S. An Overview of Electronic Commerce (E-commerce). 2021. *Journal of Contemporary Issues in Business and Government*, 27(3).

JAIN, V., WADHWANI, K., EASTMAN, J. K. 2023. Artificial intelligence consumer behavior: A hybrid review and research agenda. *Journal of Consumer Behavior*.

KHAJVAND, M., ZOLFAGHAR, K., ASHOORI, S., ALIZADEH, S. 2011. Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, 3, 57–63.

PEDREGOSA, F., VAROQUAUX, GA"EL, GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825–2830.

RAHIM, M. A., MUSHAFIQ, M., KHAN, S., ARAIN, Z. A. 2021. RFM-based repurchase behavior for customer classification and segmentation. *Journal of Retailing and Consumer Services*, 61, 102566.

ROUSSEEUW, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 53-65.

ROSÁRIO, A., RAIMUNDO, R. 2021. Consumer Marketing Strategy and E-commerce in the Last Decade: A Literature Review. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(7), 3003–3024.

SARKER, I. H. 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3), 160.

SARKER, I. H. 2022. AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems. *SN Computer Science*, 3(2), 158.

SIEBERT, M., KOHLER, C., SCERRI, A., & TSATSARONIS, G. Technical Background and Methodology for the Elsevier's Artificial Intelligence Report. 2018.

VERMA, S., SHARMA, R., DEB, S., MAITRA, D. 2021. Artificial intelligence in marketing: Systematic review and future research direction. *International Journal of Information Management Data Insights*, 1(1), 100002.

WIRTH, R., HIPPI, J. 2000, April. CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of knowledge discovery and Data Mining* (Vol. 1, pp. 29-39).

WOLD, S., ESBENSEN, K., GELADI, P. 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1–3), 37–52.

ZHOU, J., WEI, J., XU, B. 2021. Customer segmentation by web content mining. *Journal of Retailing and Consumer Services*, 61, 102588.